Journal of Combinatorial Theory, Series A **92**, 54–60 (2000)
doi:10.1006/jcta.1999.3050, available online at http://www.idealibrary.com on IDEAL®

Perfect Hash Families: Probabilistic Methods and Explicit Constructions

Simon R. Blackburn[1]

Department of Mathematics, Royal Holloway, University of London, Egham, Surrey TW20 0EX, United Kingdom
E-mail: s.blackburn@rhbnc.ac.uk

Communicated by the Managing Editors

Received June 10, 1999

An (n, q, t)-perfect hash family of size s consists of a set V of order n, a set F of order q, and a sequence of s functions from V to F with the property that for any subset X of V of order t, at least one of the functions is injective when restricted to X. An (n, q, t)-perfect hash family of minimal size is known as optimal. The paper presents a probabilistic existence result for perfect hash families which improves on the well known result of Mehlhorn for many parameter sets. The probabilistic methods are strong enough to establish the size of an optimal perfect hash family in many cases. The paper also gives several explicit constructions of classes of perfect hash families. © 2000 Academic Press

Key Words: perfect hash families; probabilistic methods.

1. INTRODUCTION

Let n, q, t, and s be positive integers and suppose (to avoid trivialities) that $n > q \geqslant t \geqslant 2$. Let V be a set of cardinality n and let F be a set of cardinality q. We say that a function $\phi: V \to F$ *separates* a subset X of V if ϕ is an injection when restricted to X; otherwise we say that ϕ *reduces* X. An (n, q, t)-*perfect hash family of size* s is a sequence $\phi_1, \phi_2, ..., \phi_s$ of functions from V to F with property that for all sets $X \subseteq V$ such that $|X| = t$, at least one of $\phi_1, \phi_2, ..., \phi_s$ separates X. The notation PHF$(s; n, q, t)$ is used for an (n, q, t)-perfect hash family of size s. A perfect hash family is *optimal* if s is as small as possible given n, q, and t.

Perfect hash families were introduced by Mehlhorn [12] in the mid-1980s. They were used in compiler design to prove lower bounds on the size of a program that constructs a hash function suitable for fast retrieval of fixed data such as library function names; see Czech, Havas, and Majewski [7] for a recent survey of this area. In the last few years,

[1] The author is an EPSRC Advanced Fellow.

0097-3165/00 $35.00

perfect hash families have been applied to circuit complexity problems (see Newman and Wigderson [13]), to the construction of deterministic analogues of probabilistic algorithms (see Alon and Naor [2] and to threshold cryptography (see Blackburn, Burmester, Desmedt, and Wild [5] and Blackburn [4]). Stinson, van Trung, and Wei [15] have used perfect hash families (and the closely related separating hash families) to improve explicit constructions of secure frameproof codes, key distribution patterns, group testing algorithms, cover-free families and separating systems. Papers of Fredman and Komlós [9], Alon [1], Körner and Marton [11], Atici, Magliveras, Stinson and Wei [3], Blackburn and Wild [6], and Stinson, Wei, and Zhu [16] consider perfect hash families from a combinatorial point-of-view.

This paper contains a probabilistic existence result for perfect hash families, that improves on the classical probabilistic methods of Mehlhorn [12] for many parameter sets. This result is given in Section 2. Moreover, this section uses a bound of Blackburn and Wild [6] to show that the result of Section 2 is the best possible in many cases. Finally, Section 3 describes several explicit constructions of good classes of perfect hash families in the case when $t=3$ or $t=4$.

2. PROBABILISTIC METHODS

A classical result due to Mehlhorn [12], proved using straightforward probabilistic methods, states that a PHF$(s; n, q, t)$ exists whenever

$$s > \frac{\log\binom{n}{t}}{\log q^t - \log\left(q^t - t!\binom{q}{t}\right)}. \tag{1}$$

We will improve this result for many parameter sets using the Lovász Sieve—see Erdős and Lovász [8] or Spencer [14]. The lemma we use may be stated as follows. Let $A_1, A_2, ..., A_r$ be events. A graph Γ on the vertex set $\{1, 2, ..., r\}$ is a *dependency graph* for the events $A_1, A_2, ..., A_r$ if for all $i \in \{1, 2, ..., r\}$ the event A_i is independent of the joint distribution of the events A_k where k is not adjacent to i in Γ.

LEMMA 1 [8, 14]. *Let Γ be a dependency graph for the events $A_1, A_2, ..., A_r$. Assume that the maximum degree of a vertex of Γ is m and assume that $\Pr(A_i) \leqslant p$ for all i. Provided that $4mp < 1$, we have that $\bigcap \overline{A_i} \neq \varnothing$.*

THEOREM 1. *A* PHF$(s; n, q, t)$ *exists whenever*

$$s > \frac{\log 4\left(\binom{n}{t} - \binom{n-t}{t}\right)}{\log q^t - \log\left(q^t - t!\binom{q}{t}\right)}. \tag{2}$$

Proof. Let s, n, q, and t be fixed. Let V be a set of cardinality n and let F be a set of cardinality q. Let $\phi_1, \phi_2, ..., \phi_s$ be functions chosen uniformly and independently at random from the set of all functions from V to F. For any subset $X \subseteq V$ such that $|X| = t$, let A_X be the event that X is reduced by all of $\phi_1, \phi_2, ..., \phi_s$. Note that $\phi_1, \phi_2, ..., \phi_s$ form a PHF$(s; n, q, t)$ if and only if none of the events A_X occur. Hence a PHF$(s; n, q, t)$ exists provided that $\bigcap \overline{A_X} \neq \varnothing$.

It is easy to see that $\Pr(A_X) = p$ where

$$p = \left(\frac{q^t - t!\binom{q}{t}}{q^t}\right)^s.$$

Let Γ be the graph whose vertices are identified with the t-subsets X of V, and where vertices identified with $X_1, X_2 \subseteq V$ are joined by an edge precisely when $X_1 \cap X_2 \neq \varnothing$. Now Γ is a dependency graph of the events A_X, and the vertices of Γ have degree m where $m = \binom{n}{t} - \binom{n-t}{t}$. By Lemma 1 a PHF$(s; n, q, t)$ exists provided that $4mp < 1$. Taking logarithms of both sides of this inequality and rearranging, we find that this condition is equivalent to the inequality (2). ∎

The bound (2) proved in Theorem 1 is better than the classical bound (1) whenever $\frac{3}{4}\binom{n}{t} < \binom{n-t}{t}$; this inequality holds when t is small when compared with n. For example, let t be a fixed integer such that $t \geqslant 2$ and let d be a fixed real number such that $d > 1$. The bound of Theorem 1 shows that a PHF$(\lfloor d(t-1) \rfloor + 1; \lfloor q^d \rfloor, q, t)$ exists for all sufficiently large integers q, whereas the classical bound only shows the existence of a PHF$(\lfloor dt \rfloor + 1; \lfloor q^d \rfloor, q, t)$. This result compares well with the best known constructions for such a class of perfect hash families—Blackburn and Wild [6] show that there exists a PHF$(d(t-1); q^d, q, t)$ whenever d is an integer and q is a sufficiently large prime power.

We now show that the result of Theorem 1 is tight for many parameters. We use the following result to be found in Blackburn and Wild [6, Theorem 1].

THEOREM 2. *Let e be an integer such that $e \geqslant 2$. Suppose a* PHF$(s; n, q, t)$ *exists with $n > (t-1)\,q^e$. Then $s \geqslant (t-1)\,e+1$.*

THEOREM 3. *Let e and t be integers such that $e, t \geqslant 2$. Let d be a real number such that $0 < d-e < 1/(t-1)$. For sufficiently large q, an optimal $(\lfloor q^d \rfloor, q, t)$-perfect hash family has size $(t-1)\,e+1$.*

Proof. Since $d-e>0$, we have that $\lfloor q^d \rfloor > (t-1)\,q^e$ for all sufficiently large q. Hence, by Theorem 2, when q is sufficiently large any PHF$((t-1)\,e+1; \lfloor q^d \rfloor, q, t)$ is optimal.

Theorem 1 asserts that a PHF$(s; n, q, t)$ exists whenever

$$s > \frac{\log 4\left(\binom{n}{t} - \binom{n-t}{t}\right)}{\log q^t - \log\left(q^t - t!\binom{q}{t}\right)}.$$

As $q \to \infty$, with $n = \lfloor q^d \rfloor$, the right hand side of this expression tends to $d(t-1)$. Since $d-e<1/(t-1)$, we have that $d(t-1) < (t-1)\,e+1$ and so Theorem 1 shows that a PHF$((t-1)\,e+1; \lfloor q^d \rfloor, q, t)$ exists whenever q is sufficiently large. ∎

3. EXPLICIT CONSTRUCTIONS

This section presents two simple constructions of classes of perfect hash families in the case when $t=3$ or $t=4$.

Let r be a fixed integer such that $r \geqslant 2$. We may construct an optimal PHF$(3; r^3, r^2, 3)$ as follows. Let T be a set of size r. Set $V = T^3$ and $F = T^2$. Define functions $\phi_1, \phi_2, \phi_3 : V \to F$ by

$$\phi_1((a, b, c)) = (a, b),$$

$$\phi_2((a, b, c)) = (b, c)$$

and

$$\phi_3((a, b, c)) = (a, c),$$

for all $a, b, c \in T$.

THEOREM 4. *The functions ϕ_1, ϕ_2, and ϕ_3 defined above form an optimal* PHF$(3; r^3, r^2, 3)$.

Proof. We need to verify the perfect hash family property. Since an element $x \in V$ is uniquely determined by any two of three images $\phi_1(x)$, $\phi_2(x)$, and $\phi_3(x)$, every 2-subset of V is reduced by at most one of ϕ_1, ϕ_2, and ϕ_3. Suppose, for a contradiction, that $\{x_1, x_2, x_3\} \subseteq V$ is a 3-set that is reduced by all of the functions ϕ_1, ϕ_2, and ϕ_3. Since every 2-set is reduced by at most one of ϕ_1, ϕ_2, and ϕ_3 and since every function ϕ_i must reduce some 2-subset of $\{x_1, x_2, x_3\}$, we may assume without loss of generality that $\phi_1(x_1) = \phi_1(x_2)$, $\phi_2(x_1) = \phi_2(x_3)$, and $\phi_3(x_2) = \phi_3(x_3)$. Writing $x_i = (a_i, b_i, c_i)$ for all $i \in \{1, 2, 3\}$, this implies that $a_1 = a_2$, $b_1 = b_2$, $b_1 = b_3$, $c_1 = c_3$, $a_2 = a_3$, and $c_2 = c_3$, and these equalities show that $x_1 = x_2 = x_3$. This contradiction establishes that ϕ_1, ϕ_2, and ϕ_3 form a perfect hash family.

Corollary 2 of Blackburn and Wild [6] implies that a PHF(s; r^3, r^2, 3) has the property that $s \geqslant 3$ provided that $r > 2$. This implies that ϕ_1, ϕ_2, ϕ_3 form an optimal perfect hash family when $r > 2$. An easy *ad hoc* argument shows that a PHF(2; n, 4, 3) must have $n \leqslant 6$, and so ϕ_1, ϕ_2, ϕ_3 form an optimal perfect hash family when $r = 2$. ∎

We may construct a PHF(6; p^2, p, 4) for any prime number p such that $p = 11$ or $p \geqslant 17$ as follows. Let $F = \mathbb{F}_p$, the finite field of order p. Let $V = F^2$. Define $\phi_1, \phi_2, ..., \phi_6$ by

$$\phi_1((a, b)) = a,$$

$$\phi_2((a, b)) = b,$$

$$\phi_3((a, b)) = b - a,$$

$$\phi_4((a, b)) = b - 2a,$$

$$\phi_5((a, b)) = b - 3a,$$

$$\phi_6((a, b)) = b - 5a$$

for all $a, b \in F$.

THEOREM 5. *The functions ϕ_i defined above form a* PHF(6; p^2, p, 4) *for all prime numbers p such that $p = 11$ or $p \geqslant 17$.*

A sketch proof of the theorem goes as follows. A set of 4 points in the plane $\mathbb{F}_p^2$ gives rise to at most 6 elements of $\mathbb{F}_p \cup \{\infty\}$, corresponding to the gradients of the lines passing through each pair of the 4 points. It is not difficult to check that a 4-set of points is reduced by all of the functions $\phi_1, \phi_2, \phi_3, \phi_4, \phi_5$, and ϕ_6 if and only if this set of gradients is $\{\infty, 0, 1, 2, 3, 5\}$. Now, if 5 of the gradients between pairs of points in the 4-set are specified, the final gradient is uniquely determined. Using this fact,

it is easy (but tedious) to check that when $p = 11$ or $p \geqslant 17$ no set of 4 points is associated with the set $\{\infty, 0, 1, 2, 3, 5\}$ of gradients. Hence every 4-set in V is separated by at least one of the functions ϕ_i, and so the functions ϕ_i form a perfect hash family with the parameters given.

We remark that this construction produces explicit examples of linear perfect hash families; see Blackburn and Wild [6]. It would be very interesting to determine whether this class of perfect hash families is optimal (at least for sufficiently large values of p). The bound given in [6, Theorem 2] shows that a PHF(s; p^2, p, 4) must satisfy the inequality $s \geqslant 5$, but no infinite classes of families of the form PHF(5; p^2, p, 4) are known.

The above construction does not work when $p = 13$—the set $\{(0, 0), (0, 1), (3, 3), (1, 3)\}$ is not separated by any of the functions $\phi_1, \phi_2, ..., \phi_6$. In fact, a small computer search shows that no PHF(6; 13^2, 13, 4) can be constructed using linear methods as above. However, it is easy to construct a PHF(7; 13^2, 13, 4) by taking any 7 distinct linear functions from $\mathbb{F}_p^2$ to $\mathbb{F}_p$. It would be interesting to determine whether a PHF(s; 13^2, 13, 4) exists when $s < 7$.

If we fix integers d and t, we may ask whether there exists a positive constant c such that there exists an infinite collection of perfect hash families of the form PHF($d(t-1)-1$; cq^d, q, t). This question is even open for $d = 2$ and $t = 3$. In this special case, a bound of Blackburn and Wild [6, Corollary 2] shows that $c \leqslant 1/2$ if it exists. The largest known example of a PHF(3; $\frac{1}{2}q^2$, q, 3) is given in Fig. 1. Here the points of V are identified with the positions of the grid that are labelled with an integer. If position (i, j) is labelled with k, then the associated point x is such that $\phi_1(x) = i$, $\phi_2(x) = j$ and $\phi_3(x) = k$. (To verify the perfect hash family property, one must verify that any rectangle with two corners labelled with the same value has its other two corners unlabelled.) There is strong evidence for the belief that this perfect hash family is an extreme example, and that $c \leqslant 1/3$ if it exists.

1			2		3	4	
	5	6		7			8
7		4			5		2
	3		8	1		6	
6			5	4			3
	2	1			8	7	
8		3		2		5	
	4		7		6		1

FIG. 1. A PHF(3; 32, 8, 3).

ACKNOWLEDGMENTS

Many thanks to Jack van Lint, for an inspiring talk on the use of the Lovász sieve in the context of codes having the Identifiable Parent Property; see the paper by Hollmann, van Lint, Linnartz, and Tolhuizen [10]. Thanks also to Peter Wild, for reading an earlier version of this paper.

REFERENCES

1. N. Alon, Explicit construction of exponential sized families of k-independent sets, *Discrete Math.* **58** (1986), 191–193.
2. N. Alon and M. Naor, Rerandomization, witnesses for Boolean matrix multiplication and construction of perfect hash functions, *Algorithmica* **16** (1996), 434–449.
3. M. Atici, S. S. Magliveras, D. R. Stinson and W.-D. Wei, Some recursive constructions for perfect hash families, *J. Combin. Des.* **4** (1996), 353–363.
4. S. R. Blackburn, Combinatorics and threshold cryptography, *in* "Combinatorial Designs and Their Applications" (F. C. Holroyd, K. A. S. Quinn, C. Rowley, and B. S. Webb, Eds.), Research Notes in Mathematics, Vol. 403, pp. 44–70, CRC Press, London, 1999.
5. S. R. Blackburn, M. Burmester, Y. Desmedt, and P. R. Wild, Efficient multiplicative sharing schemes, *in* "Advances in Cryptology—EUROCRYPT '96" (U. Maurer, Ed.), Lecture Notes in Computer Science, Vol. 1070, pp. 107–118, Springer-Verlag, Berlin, 1996.
6. S. R. Blackburn and P. R. Wild, Optimal linear perfect hash families, *J. Combin. Theory Ser. A* **83** (1998), 233–250.
7. Z. J. Czech, G. Havas, and B. S. Majewski, Perfect hashing, *Theoret. Comput. Sci.* **182** (1997), 1–143.
8. P. Erdős and L. Lovász, Problems and results on 3-chromatic hypergraphs and some related questions, *in* "Infinite and Finite Sets" (A. Hajnal, R. Rado, and V. Sós, Eds.), Colloquium Math. Soc. Janos Bolyai, Vol. 11, pp. 609–627, North-Holland, Amsterdam, 1975.
9. M. L. Fredman and J. Komlós, On the size of separating systems and families of perfect hash functions, *SIAM J. Alg. Discrete Methods* **5** (1984), 61–68.
10. H. D. L. Hollmann, J. H. van Lint, J.-P. Linnartz, and L. M. G. M. Tolhuizen, On codes with the identifiable parent property, *J. Combin. Theory Ser. A* **82** (1998), 121–133.
11. J. Kőrner and Marton, New bounds for perfect hashing via information theory, *European J. Combin.* **9** (1988), 523–530.
12. K. Mehlhorn, "Data Structures and Algorithms. 1. Sorting and Searching," Springer-Verlag, Berlin, 1984.
13. I. Newman and A. Wigderson, Lower bounds on formula size of Boolean functions using hypergraph entropy, *SIAM J. Discrete Math.* **8** (1995), 536–542.
14. J. Spencer, Probabilistic methods, *Graphs Combin.* **1** (1985), 357–382.
15. D. R. Stinson, T. van Trung, and R. Wei, Secure frameproof codes, key distribution patterns, group testing algorithms and related structures, *J. Statist. Plann. Inference*, in press.
16. D. R. Stinson, R. Wei, and L. Zhu, New constructions for perfect hash families and related structures using combinatorial designs and codes, *J. Combin. Designs*, to appear.